Application of GIS and logistic regression to fossil pollen data in modelling present and past spatial distribution: A case study in the Colombian Savanna

S.G.A. Flantua, J.H. van Boxel, H. Hooghiemstra, J. van Smaalen
Institute for Biodiversity and Ecosystem Dynamics, Universiteit van Amsterdam,
Kruislaan 318, 1098 SM Amsterdam, The Netherlands
Tel. +31-20 525 6216; Fax +31-20 525 7832
Email: J.H.Boxel@science.uva.nl

1. INTRODUCTION

Climate change at glacial-interglacial cycle time scales have had an impact on the vegetation in many parts of the world. Vegetation change is reflected by changes in abundance, geographic extent and floral composition of plant populations, from which pollen grains are naturally preserved in lakes and peat bogs. By drilling at sites of interest, sediment cores are obtained which show the temporal variations in the pollen assemblages. Palynologists present these data in pollen diagrams and interpret the downcore changes in pollen spectra in terms of past vegetation change and inferred environmental conditions.

So far little research has been carried out, in which palynological data is analysed by software specially designed for spatial data analysis, like Geographical Information Systems (GIS), although the arguments for its implementation are divers: palynological datasets in general are large and complex to interpret; the data consists of changes which have occurred over an area (2-dimensional surface) and over time (a third dimension-variable); and frequently data from different locations must be compared to make an interpretation of a complete area rather than of one single site only.

The implementation of GIS in palynological research appears to be in an explorative stage as the examples are scarce, e.g.: displaying plant-distributions (e.g. Giesecke and Bennett, 2004), habitat suitability analysis (e.g. Lyford *et al.*, 2003) and the reconstruction of past vegetation with historical maps (e.g. Veski *et al.,* 2005). Due to the complexity and the spatial heterogeneity of the variables influencing the spatial distribution of vegetation, palynological analysis thus far seems to be limited by non-spatial methods, hence trying to find structures in a multidimensional data set with one-dimensional tools.

The aim of this paper is to provide a useful step in pollen data analysis by merging GIS and predictive modelling into a combined palynological GIS application. The proposed methodology can be employed by palynologists to explore their area of research and capture it in a predictive model, to make reconstructions of past and future land-cover distributions under changing climatic conditions. This provides a better understanding of the vegetation responses to alternating environmental conditions which subsequently contributes to the interpretation of the pollen data. In the same database the palynological data can be implemented to use GIS to reconstruct and evaluate patterns of land-cover changes based on pollen counts.

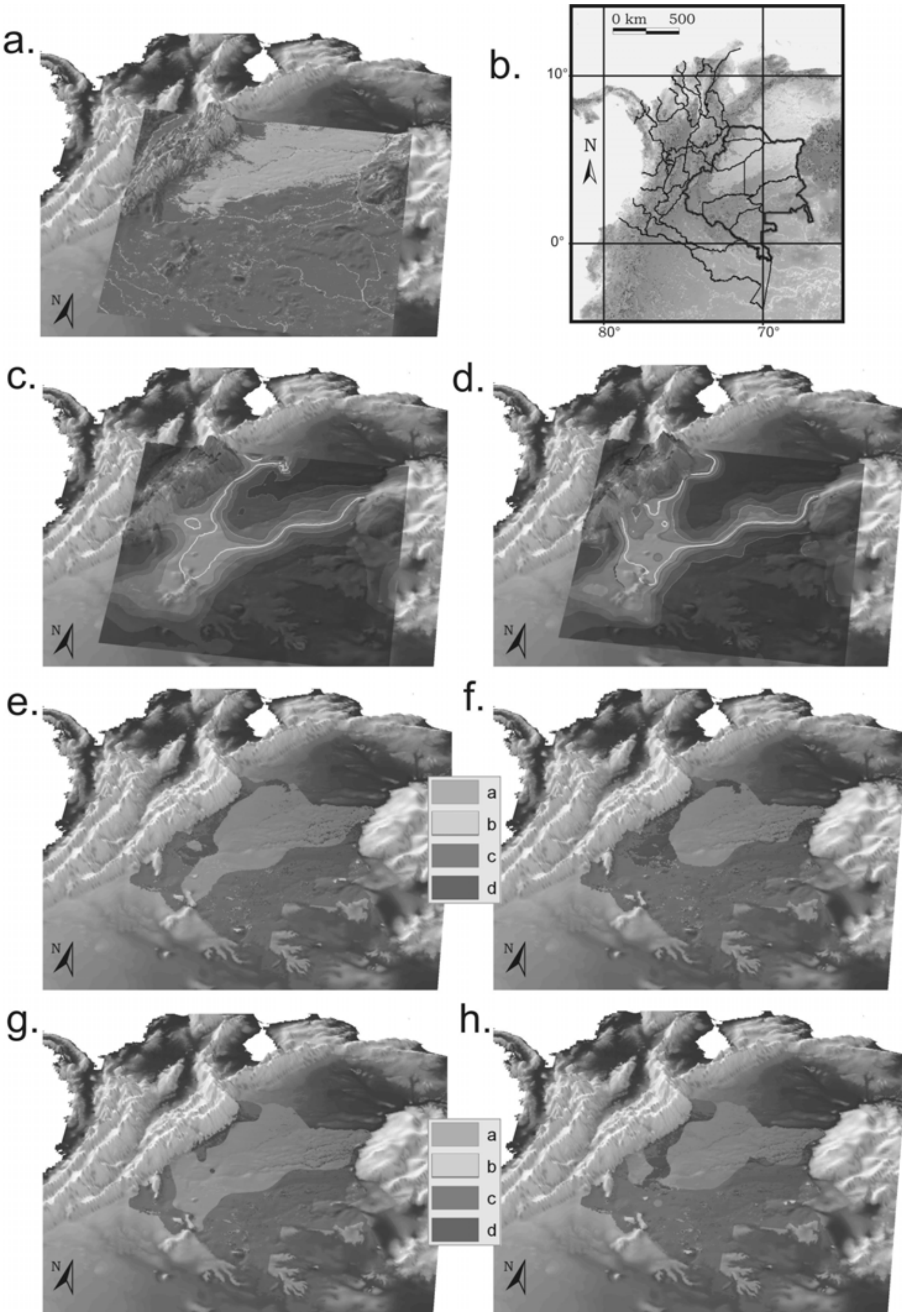

Figure 1: *Study area and outcomes of predicted land-cover distribution by logistic model. (a) Map of South-America indicating the location of the study area and the actual land-cover distribution; (b) Location of the Colombian savanna biome in between the Andes and the Guyana Shield (03 to 07 ºN, 68 to 71 ºW); (c) Probability map of savanna occurrence based on random data sampling; (d) Probability map of savanna occurrence based on regular sampling. Yellow lines indicate the 0.5 threshold whereas the red lines delineate the 0.6 cut-point; (e-h) Differences between actual land-cover distribution and predicted by the model; (e) Based on random sampling at 0.5 threshold; (f) Random sampling at 0.6 threshold; (g) Regular sampling at 0.5 threshold; (h) Regular sampling at 0.6 threshold. Legend-specification: (a) Indicates correctly predicted savanna [Dark yellow] and (b) represents where the model falsely predicted savanna [Bright green] (c) Indicates correctly predicted forest [Dark green], while (d) shows where the model failed to predict savanna [Red].*

To illustrate the implementation of the palynological GIS application, an area of palynological research is selected in the tropical lowlands of northern South-America. Colombia and Venezuela share an extended area of savannas, which expands from the Eastern Andes Cordillera all the way to the eastern coast of Venezuela (Fig. 1a,b). The southern boundary of the Colombian savanna which is transitional to tropical rainforest, has migrated through time, as these geographical shifts have been reconstructed from pollen records close to this zone (Behling and Hooghiemstra, 1998). These changes indicate that a certain level of vegetation dynamics has occurred, but the degree of environmental change is mostly expressed in general terms, such as "to some extent drier or wetter" conditions, or lacking specification regarding the seasonality, like "shorter" or "longer dry period". By using the palynological GIS application as analytic tool, a better understanding of the dynamics of the savanna distribution in Colombia can be realized, while furthermore contributing to a better understanding of future vegetation responses to global climate change.

The employment of GIS is basically divided into two different but related applications. The first application is to make a predictive model, in which the climatic variables are determined which influence the spatial distribution of the savanna vegetation in our area of the interest. The statistical model, derived from logistic regression, is subsequently implemented in GIS and used to create land-cover maps, which are compared to the actual land-cover distribution. The models deviations are made clear by spatial analysis of the predictor variables to assess the vegetation response to the climate. The second application introduces data from pollen records of the Colombian savanna into GIS to create land-cover maps through pollen percentages implementation and interpolation methods. An assessment is made of the suitability of the pollen data for a GIS analysis in which both limitations and recommendations are discussed (for more detailed results see in Flantua et al. 2007).

2. METHODOLOGY

Logistic regression is a variation of ordinary regression, which basically is a method used to determine the impact of independent variables on a dependent variable. In binary logistical regression the outcome is restricted to two values, representing the presence or absence of a specific event. It produces a formula that predicts the probability of the occurrence as a function of the independent variables. Depending on a chosen threshold probability value, everything above this threshold indicates one condition of the binomial outcome (i.e. the presence of savanna), while everything below equals the other condition (i.e. absence of savanna; in this case presence of forest).

A binomial land-cover GIS layer is created from the Global Land-cover Characteristics (GLCC) Data Base Version 2.0 (Loveland *et al.*, 2000). The monthly values of precipitation, temperature and potential evapotranspiration form the basis for the set of independent (predictor) variables in the model creation (Legates and Willmott, 1990a,b). By using a point layer raster that overlies all data layers in GIS, the underlying variables values at each point can be extracted and readily used for the statistical analysis and model creation. The resulting statistical formula of variables is introduced into GIS with layers of the most significant climate predictors as data source in order to get a map of the predicted land-cover distribution as result. This map is subsequently compared to the actually observed spatial distribution of savanna to visualize the areas in which the climatic conditions fall short in correctly predicting the vegetation pattern. Different sampling methods, threshold values and accuracy assessments methods are used to evaluate the difference in predictive capacity. Implementing the logistic regression outcome into GIS does not only test the models predictive capacity on a larger data set, but also provides insight into a vegetation distribution based on only climate variables, and therefore the models weaknesses.

To obtain a reconstruction of temporal land-cover changes in the past, pollen spectra at successive time slices are compared. The time slices of interest were selected based on the amount of available pollen data and the degree of change compared to earlier time slices to make meaningful intervals. Two different interpolation methods are used to create the land-cover reconstructions: Local Polynomial and Radial Basis Functions.

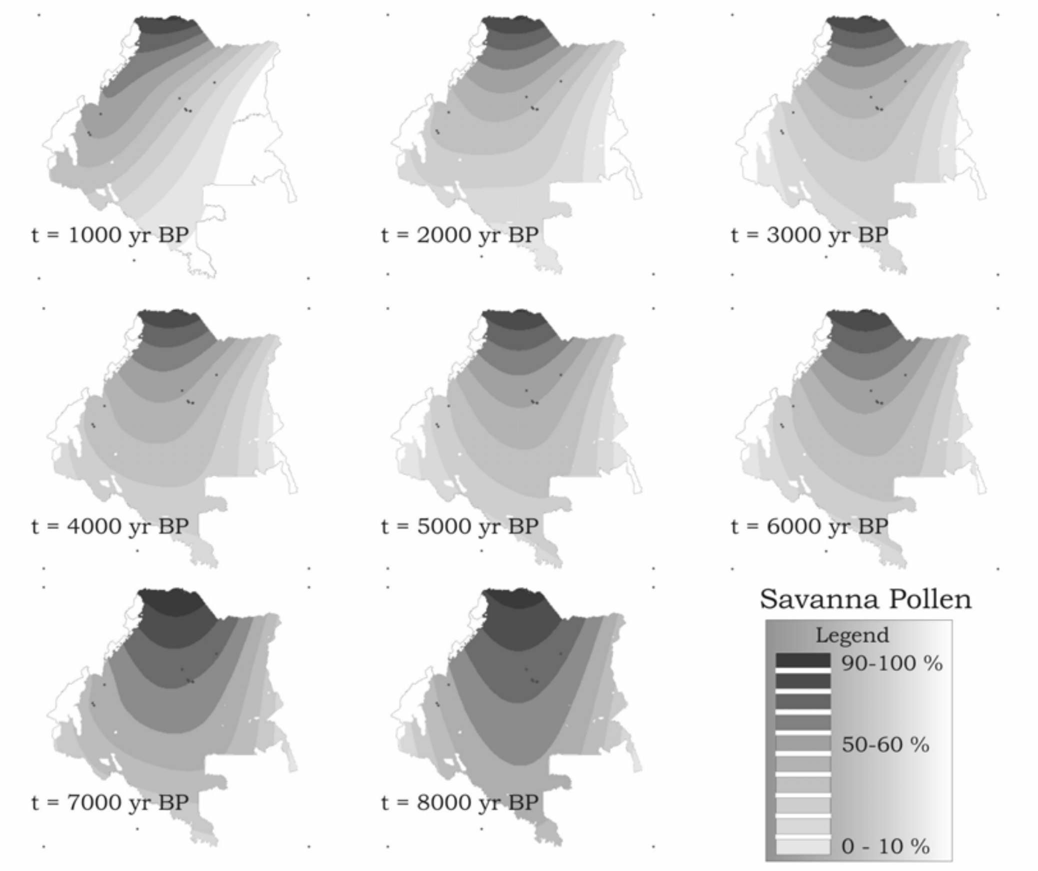

Figure 2: *Maps of interpolated pollen percentages of taxa reflecting savanna vegetation based on Local Polynomial Interpolation method. The interpolated area corresponds to the area delineated in Fig. 1b. Selected time slices range from 1000 to 8000 14C yr BP. (Interpolation specification: power = 2, Ideal weight distance activated)*

3. RESULTS

The variables indicated by the predictive model as determinants of the savanna distribution, correspond to earlier publications (Sarmiento, 1983; San Jose *et al*., 1998; Rippstein *et al*., 2001; Hooghiemstra *et al*., 2002). The spatial correlation and separate influence of the variables was tested as they were independently removed from the model, to evaluate its significant contribution to the model predictive accuracy. Based on the different accuracy measurements, all created models achieve an acceptable predictive power ranging from an overall accuracy of 0.81 to 0.86 (Fig. 1c-h). A regular sampling methodology (Fig. 1d, f, h) and a 0.6 threshold value (red line in Fig.1c, d) give more robust results than random point sampling (Fig. 1c, e, g) and the default 0.5 threshold (yellow line in Fig.1c, d) achieving combined a best fit model of 85,7% correctly predicted land-cover distribution.

This difficulty of interpreting the interpolated maps (Fig. 2 and 3) is caused by different factors, including the chosen interpolation method and the pollen transect orientation. Selecting a proper method basically means a trial-and-error application to see which method is best used on the specific dataset. The pollen transect fails to represent the past land-cover shifts, due to the dissimilar orientation compared to the occurred geographical shifts, and the presence of gallery forest which significantly prevents the pollen income of possibly dominant savanna land-cover.

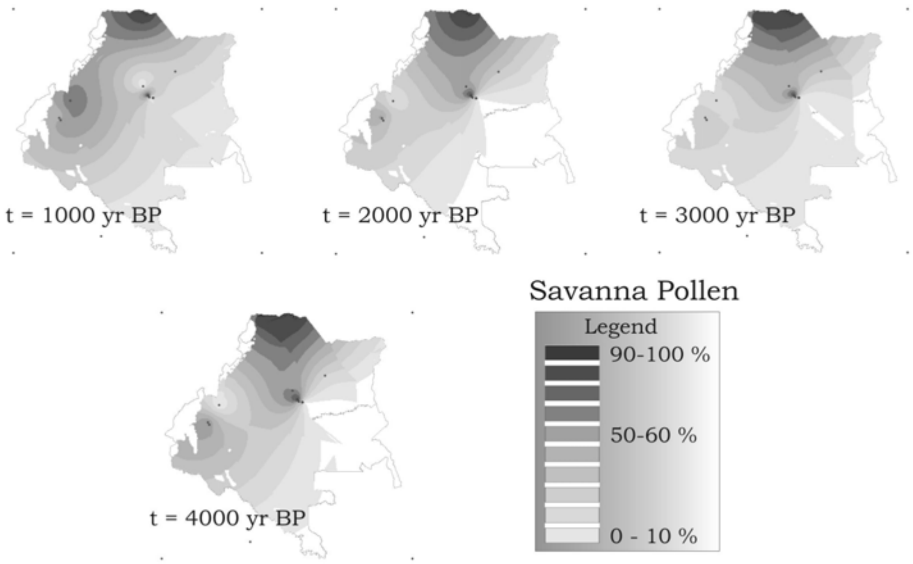

Figure 3: *Maps of interpolated pollen percentages of taxa reflecting savanna vegetation based on Radial Based Functions interpolation method. The interpolated area corresponds to the area delineated in Fig. 1b. Selected time slices range from 1000 to 4000 14C yr BP.*

4. ANALYSIS

By implementing the logistic regression model into GIS, the weaknesses of the model become evident. It is in this part of the ecological modelling that the usefulness of GIS is shown, seeing that the spatial patterns of the models are directly compared to the true land-cover patterns. Not only can the interpretation of the pattern of errors, contribute to an improvement of the model, but far more to an understanding of the responsiveness of the land-cover to different environmental conditions and therefore to the system as a whole. By a spatial analysis in GIS of this Colombian savanna model, the deviating predictions due to the influence of the total annual precipitation, the fire occurrence and human interference in the natural system are shown.

Comparing the resulting maps of different interpolations methods used (Fig. 2 and 3) shows that locations of pollen sites should be more evenly distributed over the study area to be able to improve the understanding of the geographical migration of land-cover boundaries in space and time with the help of GIS interpolation methods. Linear transects are better suitable for research areas with an altitudinal gradient where vegetation is migrating along slopes. The relevance of information on the modern pollen rain to understand the local conditions of a palynological research site, becomes strongly evident.

The combination of GIS with palynological data can further employed: to improve of site-specific information (e.g. matching estimated coordinates with satellite images), to locate potential drilling locations and to enhance the visual presentation of research proposals and/or results.

5. REFERENCES

Behling H, Hooghiemstra H (1998) Late Quaternary palaeoecology and palaeoclimatology from pollen records of the savannas of the Llanos Orientales in Colombia. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 139: 251-267

Giesecke T, Bennett KD (2004), The Holocene spread of *Picea abies* (L.) Karst. in Fennoscandia and adjacent areas. *J. Biogeogr.* 31: 1523-1548

Flantua SGA, Van Boxel JH, Hooghiemstra H, Van Smaalen J (2007) Application of GIS and logistic regression to fossil pollen data in modelling present and past spatial distribution of the Colombian savanna. Submitted to *Clim. Dynam.* (in review)

Hooghiemstra H, Van der Hammen T, Cleef A (2002) Evolution of forests in the Northern Andes and Amazonian lowlands during the Tertiary and Quaternary. In: MR Guariguata, GH Kattan, eds. (2002) *Ecología y conservación de bosques neotropicales.* Edictiones Libro Universitario Regional, Cartago, Costa Rica, pp 43-58

Legates DR, Willmott CJ (1990a) Mean seasonal and spatial variability in Gauge-corrected, global precipitation. *Int. J. Climatol.* 10: 111-127

Legates DR, Willmott CJ (1990b) Mean seasonal and spatial variability in global surface air temperature. *Theor. Appl. Climatol.* 41: 11-21

Loveland TR, Reed BC, Brown JF, Ohlen DO, Zhu J, Yang L, Merchant JW (2000) Development of a Global Land-cover Characteristics Database and IGBP DISCover from 1-km AVHRR Data: *Int. J. Remote Sens.* 21: 1303-1330

Lyford ME, Jackson ST, Betancourt JL, Gray ST (2003) Influence of landscape structure and climate variability on a late Holocene plant migration. *Ecol. Monogr.* 73: 567-583

Rippstein G, Escobar G, Motta F (2001) *Agroecología y biodiversidad de las Sabanas en Llanos orientales de Colombia.* Cali: Centro internacional de Agricultura Tropical (CIAT). Publication CIAT No. 322

San Jose JJ, Montes R, Mazorra M (1998) The nature of savanna heterogeneity in the Orinoco Basin. *Global Ecol. Biogeogr. Lett.* 7: 441-445

Sarmiento G (1983) The savannas of tropical America. *In*: F Bourlière (ed.), *Tropical Savannas*. Amsterdam: Elsevier, Ecosystems of the World, 13: 245-288

Veski S, Koppel K, Poska A (2005) Integrated palaeoecological and historical data in the service of fine-resolution land use and ecological change assessment during the last 1000 years in Rõuge, southern Estonia. *J. Biogeogr.* 32: 1473-1488